第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·网络与通信· 文章编号：1000—3428(2007)18—0109—03 文献标识码：A 中图分类号：TP393

集群服务器 IPv4/IPv6 接入机制的设计与实现

官尚元，薛正华，石卫强，董小社
（西安交通大学电子与信息工程学院，西安 710049）

摘 要：设计了一种将集群服务器 NVS(network virtual server)接入到 IPv4/IPv6 环境中的机制，通过将协议转换内置到 NVS 接口机的内核中以提高协议处理速度，实现应用对底层具体协议的透明使用，原有高可用性、高可扩展性和负载均衡等多接口机软件不加修改即可使用。实验结果表明，该机制协议处理延迟小于通过 NAT-PT 网关实现的延迟，同时具有更高的吞吐量。

关键词：集群服务器；IPv4；IPv6；NAT-PT

Design and Implementation of the Mechanism to Process IPv4/IPv6 Request Packets for Cluster Server

GUAN Shang-yuan, XUE Zheng-hua, SHI Wei-qiang, DONG Xiao-she
(School of Electronic & Information Engineering, Xi′an Jiaotong University, Xi′an 710049)

【Abstract】This paper presents a mechanism to deal with IPv4/IPv6 packets, which improves the speed of protocol processing through placing the module of protocol translation into the kernel of the frontal subsystem of cluster server NVS (network virtual server). This mechanism achieves that transparent applications of the underlying IP protocols and with high availability, high scalability, load balancing software of the frontal subsystem can be run on NVS without any modification. Experimental result shows that the delay of this mechanism is less than that through a NAT-PT gateway and it is characterized by high throughput.

【Key words】cluster server; IPv4; IPv6; NAT-PT

因为IPv6 相对IPv4 具有地址空间大、路由速度快、更好地支持流媒体等新型网络应用以及安全性好等优势，所以IPv6 必将取代IPv4 成为主流的网际协议。目前，NVS只能接入到IPv4 网络中[1]，故必须增加NVS对IPv6 的支持。

因特网规模巨大，更换网际协议是一个庞大的工程，不可能在短时间内完成。因此，IPv4 和IPv6 将在一段时间内共存于同一环境中。目前主要有 3 种技术解决IPv4 与IPv6 网络共存问题[1,2]：双协议栈技术(RFC2893 obsolete RFC1933)，主机或路由器在同一接口上运行IPv4/IPv6 双协议栈，根据需要选择协议栈；隧道技术(RFC2893)，将IPv6 包当作数据，穿越IPv4 网络；协议转换技术 (NAT-PT，RFC2766)，通过转换网关完成IPv6 包和IPv4 包之间的相互转换，从而实现纯IPv4 和IPv6 主机之间的透明通信。

目前，主流的操作系统(如Linux, Windows)都支持IPv4/IPv6 双协议栈，所以采用双协议栈技术易于实现，但需要NVS系统软件和应用软件能够支持IPv6，故采用该技术不利于NVS的推广和对已有系统和应用软件可移植性支持。隧道技术也有类似的缺点，而且需要对IPv6 包进行封装和解包，效率较低。采用NAT-PT技术的一种简单的方法是：在NVS和IPv6 网络之间安装一个支持IPv4 和IPv6 包相互转换的设备[3]，这种方式附加了额外设备，同时容易成为NVS的瓶颈，扩展性差、可用性低。

本文设计并实现了一种接入到IPv4/IPv6 环境中的机制。该机制能使集群系统中已实现的对等多接口机系统[4, 5]可用性、可扩展性和负载均衡等软件不加修改即可使用，同时具有协议处理延迟小等特点。

1 IPv4/IPv6 接入机制原理

为了克服基于主机的软件接口机的可扩展性差、容易成为系统瓶颈等缺点[6]，文献[5]设计实现了一种基于交换机链路聚集技术的多接口机方案，拓扑结构如图 1 所示。方案中接口机的数目能够随着服务器系统规模而增减，从而能有效增强服务器系统的I/O 能力，同时基于链路聚集特性开发了配套的高可用性，高扩展性和负载均衡等系统软件。

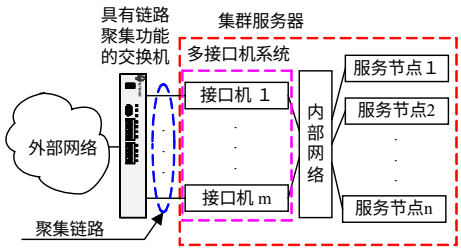


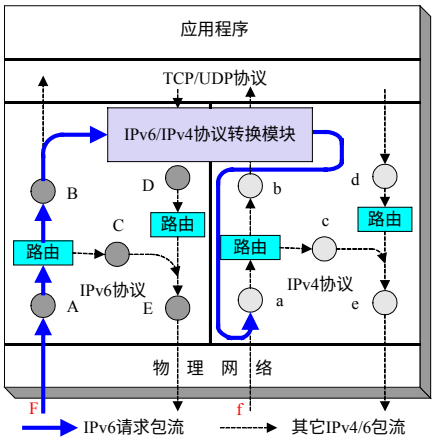
图 1 基于链路聚集的多接口机系统

为了让上层应用软件和已经开发的多接口机系统软件对外部网络透明使用，采用 NAT-PT 技术实现 IPv4/IPv6 接入机制。目前，利用 NAT-PT 实现 IPv4 和 IPv6 主机互通的协议转换软件一般运行在用户空间，为了缩减对数据包的处理流

基金项目：国家“863”计划基金资助项目(2004AA111110)
作者简介：官尚元(1979－)，男，博士研究生，主研方向：高性能服务器；薛正华，博士研究生；石卫强，硕士研究生；董小社，博士生导师、教授
收稿日期：2006-10-13 **E-mail：**ranger_79@163.com

程，将 NAT-PT 内置到接口机内核中，并通过 Netfilte 实现，这样可以提高协议处理速度。

IPv4/IPv6 接入机制处理 IPv6 请求的主要流程如下：先在接口机中将 IPv6 请求包转换成 IPv4 请求包，再由 IPv4 协议栈像其它 IPv4 包一样进行处理，如图 2 所示；将响应 IPv6 请求的 IPv4 数据包在接口机或服务节点上转换成 IPv6 包并转发到 IPv6 网络上。IPv4 请求包的处理跟未实现该接入机制时一样，从而达到对 IPv4/IPv6 包的处理。



A.NF_IP6_PRE_ROUTING:数据包在抵达路由之前经过
B.NF_IP6_LOCAL_IN:目的地为本地主机的数据包经过
C.NF_IP6_FORWARD:目的地非本地主机的数据包经过
D.NF_IP6_LOCAL_OUT:本地主机发出的数据包经过
E.NF_IP6_POST_ROUTING:数据包在离开本地主机之前经过
a,b,c,d,e含义同A,B,C,D,E

图 2　IPv6 请求包在单台接口机中的处理流程

当一个请求包到达时，由链路聚集技术和已开发的配套软件确保有且只有一台接口机接受并处理该请求。因此，该接入机制的关键在于接口机内核如何处理 IPv6 请求包，其流程如图 2 所示：IPv6 请求包从 F 处进入接口机系统，经过 IP 校验后，由 A 点的钩子函数进行处理；经过路由查找以后，送往本机的数据包经过 B 点，在 B 点判断 IPv6 地址，如果包的目标地址是 NVS，则由 IPv4/IPv6 协议转换模块进行转换并将新包插入到 IPv4 栈的等待队列中等待处理，否则交由 IPv6 协议栈继续处理。这样，一个 IPv6 请求包就转换成一个 IPv4 请求包，并像其他 IPv4 包一样处理。

响应 IPv6 请求的 IPv4 数据包也需要进行协议转换。IP 虚拟服务器软件 IPVS(IP Virtual Server)包含 3 种调度技术：通过网络地址转换实现虚拟服务器的 VS/NAT 技术，通过 IP 隧道实现虚拟服务器的 VS/TUN 技术和通过直接路由实现虚拟服务器的 VS/DR 技术。根据不同的调度技术，在接口机或服务节点上将响应数据包转换为 IPv6 格式，其原则是：如果响应包通过接口机返回给客户，则在接口机上转换；如果响应包在服务节点上直接返回，则在服务节点上转换。

当在接口机上转换时，一个响应 IPv4 包从图 3 中 f 处进入接口机系统，经过 IP 校验后，到达 a 点，在 a 点根据其目标 IPv4 地址是否为 IPv6 映射的地址来判断是否需要协议转换，如果需要，则由 IPv4/IPv6 协议转换模块进行转换并将新 IPv6 包路由并转发出去；否则交由 IPv4 协议栈继续处理。在服务节点上转换的情况与前面类似，但需要在服务节点上装载 IPv4/IPv6 协议转换模块。从服务节点发出的 IPv4 包都经过 d 点(如图 3)，于是在 d 点判断是否需要协议转换，如果需要，则调用协议转换模块进行转换，并把新 IPv6 包路由并转发出去；否则，由 IPv4 协议栈继续处理。

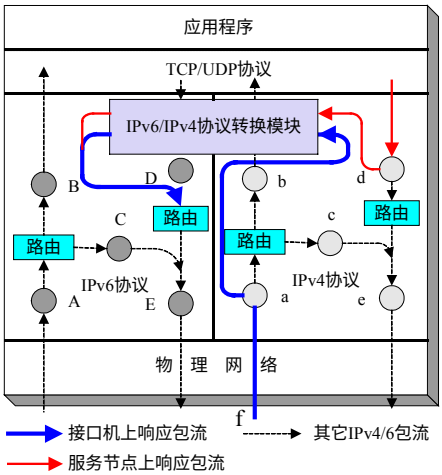


图 3　响应包在接口机或服务节点上的处理流程

通过将 IPv6 请求包转换成 IPv4 包再由接口系统处理，并将响应 IPv6 请求的数据包转换成 IPv6 包再发送到 IPv6 网络的方式，使得接口机系统软件和上层应用软件对接入环境透明使用；将协议转换模块以可加载的形式置入到接口机的内核中可提高协议转换的处理速度。

2　设计与实现

IPv4/IPv6 接入机制实现的关键在于IPv4/IPv6 协议转换模块。协议转换模块由 3 部分组成：地址/端口转换子模块，协议转换子模块和应用层转换子模块。地址/端口转换子模块维护一个IPv4 和IPv6 地址/端口映射表，完成地址/端口的相互转换；协议转换子模块完成IPv6和IPv4包之间的相互转换；应用层转换子模块负责对负载中包含IP地址信息的应用进行转换[7]。模块间相互关系以及包经过协议转换模块的处理流程如图 4 所示。

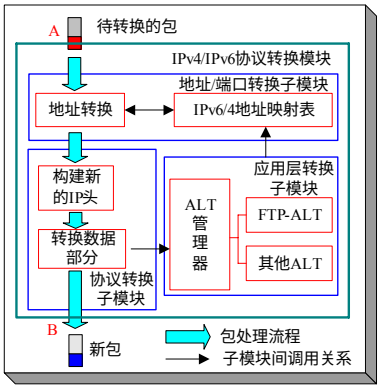


图 4　包转换处理流程以及子模块间相互关系

一个待转换的包从图 4 中的 A 处进入模块，先由地址/端口转换子模块查找 IPv6/4 地址映射表并转换地址，如果地址映射表中没有相应的映射项，则通过预设算法从地址池中分配一个地址并将该地址映射插入到表中；进入协议转换子模块后，先构建新的 IP 头，然后转换数据部分，对一般的数据，直接拷贝，如果数据是包含 IP 地址信息的应用，则需要调用应用层转换子模块进行转换，如 FTP。在完成 IP 头和数据部分的转换以后，就产生了一个新的 IP 包。

2.1　地址/端口转换子模块

该子模块主要实现 IPv4 和 IPv6 地址与端口的相互映射。有两种实现方法：一种是无状态的地址转换(stateless address translation)，它只能使用特殊的 IPv6 地址；另一种是有状态

的地址转换(stateful address translation)，它可以使用任何 IPv6 地址。后者维护一张 IPv4/IPv6 地址映射表，通过查找地址映射表来完成地址的转换，包括 3 个过程：地址邦定(address binding)、地址查找与转换(address lookup and translation)和地址解绑(address unbinding)。本文采用有状态的地址转换方法并利用动态地址映射来实现地址邦定。动态地址映射是当 IPv6 节点需要通信时，该子模块从地址池中分配一个 IPv4 地址，进行邦定，通信结束时便进行地址解绑，回收地址，以便再利用，这样提高了 IPv4 地址的利用率。由于 IPv6 和 IPv4 地址之间采用一一映射，因此无须考虑端口的转换，降低了该子模块实现难度。

在 NVS 中，接口机系统和后端服务节点构成一个内部网络，所以可以采用专为本地网使用的 10.X.X.X、172.16.X.X 和 192.168.X.X。为了方便且不失一般性，以 10.X.X.X 为例。对 8～31 位的含义进行重新规定，如图 5 所示。

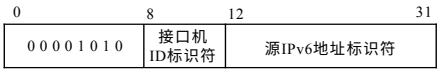


图 5　IPv6 地址 8~31 位含义

在图 5 中，0～7：标识该地址是一个临时用来映射 IPv6 地址的 IPv4 地址，以 10.开头；8～11：接口机的 ID 值，与 IPVS 中分配的 ID 标识一样，主要用来区分不同接口机，否则，不同的接口机可能会把不同的源 IPv6 地址映射成同一台 IPv4 地址从而造成混乱。目前，接口机最多 16 个，所以用 4 位表示接口机的 ID；12～31：标识源 IPv6 地址，通过某种算法对源 IPv6 地址进行操作而得到，共 20 位。

具有相同接口机 ID 的待用 IPv4 地址构成该接口机的地址池。为了提高映射表的性能，在数据结构和算法设计上采用 hash 表、链表和 LRU(最近最少使用算法)相结合的方法。映射表需要在别的接口机上备份，以配合现有高可用软件，当该接口机失效以后别的接口机接管其工作。有关映射表的建立、维护以及映射过程不在此做进一步的介绍。

2.2　协议转换子模块

该子模块主要实现 IPv4 与 IPv6 格式的相互转换。它将一个版本的包映射为另一个版本的包，包括包头转换、ICMP 转换和 TCP/UDP 转换，主要转换流程如图 6 所示。

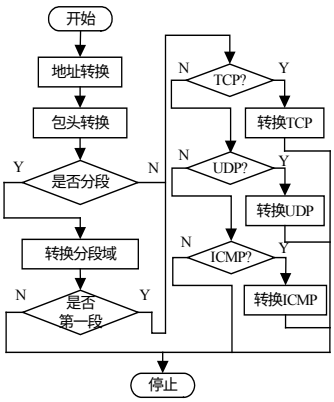


图 6　协议转换子模块转换流程

(1)包头转换：IPv4 与 IPv6 包头相似，但二者有些域不同或者含义不同或者长度不同，因此在转换时需要对不同的域进行不同的处理：有些域具有相同的大小和含义，转换时可直接复制；有些域在转换的时候值发生了变化，需要重新计算；有些域则需要被忽略；有些域无对应的语意关系，所以需要设置成静态默认值。

(2)ICMP 转换：转换器对单跳和未知类型的 ICMP 消息直接抛弃，而其余的消息具有相同的头部，所以一般可直接复制。但需要考虑如下几个方面：ICMPv6 的校验和中包括 IPv6 地址，而 ICMPv4 的校验和仅仅由 ICMP 包头决定的，所以需要重新计算校验和的值；在一些 ICMP 错误消息中，包含了导致出错的 IP 包头，而 IPv4 和 IPv6 包头大小不相同，所以需要调整指针字段的值；在 ICMP 错误消息中的 IP 数据包像普通的 IP 数据包一样需要转换，这种递归的方式将导致数据包长度发生变化。

(3)TCP/UDP 转换：因为 TCP/UDP 包含 IP 地址的伪报头，所以需要考虑校验和域。在静态转换时，由于使用特殊的 IPv6 地址且不会改变校验和，因此无须考虑校验和域。动态转换使用任意的 IPv6 地址，需要对该域进行调整。可通过计算二者之间的差值实现，计算公式如下：新校验和=旧校验和+(新分组源目地址的校验和-旧分组源目地址的校验和)。

2.3　应用层转换子模块

协议转换本身不包含转换数据包的负载部分，因此对负载中携带了 IP 地址的应用无能为力。应用层转换(ALT)就是为了解决这种情况而采用的一种机制，对负载中包含 IP 地址的典型应用进行转换，比如，IPv4 和 IPv6 中不同的 FTP 命令进行转换。把不同的 ALT 以可加载模块的形式加载到内核中并通过 ALT 管理器进行管理，这样增加了可扩展性。当有新的应用需要支持时，只需要开发相应的 ALT 并加载到内核中即可；如果不需要，则可以卸载对应的模块。不同应用的 ALT 和协议转换联合使用可对多种应用提供支持。目前，只是实现了 FTP。

由于在NVS中所有的请求都是IPv6 主机发起的，因此不需要解决由IPv4 主机发起访问IPv6 主机的请求所引起的问题[7]。

地址/端口转换子模块、协议转换子模块和应用转换子模块配合完成 IPv4 包和 IPv6 包之间的相互转换，并将 IPv4/IPv6 协议转换模块内置到接口机内核中，从而实现 NVS 接入到 IPv4/IPv6 环境中，使得应用软件和多接口机系统软件对接入环境透明使用。

3　实验与分析

主要测试两种协议转换方案的通信延迟：一种是在 NVS 和 IPv6 网络之间配置一个支持 IPv4 和 IPv6 包相互转换的网关，具体实现见文献[3]；另一种是本文实现的将协议转换置于 NVS 接口机内核中的方式。为了便于分析比较，测试一对连在同一个局域网上的纯 IPv4 主机之间的通信延迟以及连在同一个局域网上的一对纯 IPv6 主机之间的通信延迟。

利用 PING 工具测量主机之间的通信延迟。在 100Mb/s 网络上，测量 64B 到 1 024B 的 PING 数据包的往返延迟，测试结果如图 7 所示。其中，v4-v4 和 v6-v6 表示采用相同 IP 协议连在同一个局域网上的两台主机之间的延迟；GATE 表示通过转换网关互联的 IPv6 主机和 NVS 之间的延迟；PT 表示通过内置协议转换后 IPv6 主机和 NVS 之间的延迟。PT 的延迟不到 v4-v4 或 v6-v6 的 2 倍，随着包变大，延迟有所增加。不难发现，通过内置协议转换实现的延迟比通过网关实现快接近 30ms，这主要是因为将协议转换内置到内核中减少了数据包的处理流程。因为在轻负载时接口机的个数不会影响数据包的通信延迟，所以在实验中只配置了一台接口机。

(下转第 127 页)